Communications for Statistical Applications and Methods
2018, Vol. 25, No. 2, 199–215

https://doi.org/10.29220/CSAM.2018.25.2.199
Print ISSN 2287-7843 / Online ISSN 2383-4757

Forecasting daily PM_{10} concentrations in Seoul using various data mining techniques

Ji-Eun Choi[a], Hyesun Lee[a], Jongwoo Song[1,a]

[a]Department of Statistics, Ewha Womans University, Korea

Abstract

Interest in PM_{10} concentrations have increased greatly in Korea due to recent increases in air pollution levels. Therefore, we consider a forecasting model for next day PM_{10} concentration based on the principal elements of air pollution, weather information and Beijing $PM_{2.5}$. If we can forecast the next day PM_{10} concentration level accurately, we believe that this forecasting can be useful for policy makers and public. This paper is intended to help forecast a daily mean PM_{10}, a daily max PM_{10} and four stages of PM_{10} provided by the Ministry of Environment using various data mining techniques. We use seven models to forecast the daily PM_{10}, which include five regression models (linear regression, Randomforest, gradient boosting, support vector machine, neural network), and two time series models (ARIMA, ARFIMA). As a result, the linear regression model performs the best in the PM_{10} concentration forecast and the linear regression and Randomforest model performs the best in the PM_{10} class forecast. The results also indicate that the PM_{10} in Seoul is influenced by Beijing $PM_{2.5}$ and air pollution from power stations in the west coast.

Keywords: PM_{10} concentration, linear regression, Randomforest, gradient boosting, support vector machine, neural network, ARFIMA

1. Introduction

Recently, Korea scored 45.51 out of 100 in the "Environmental Performance Index 2016" in air quality, which was announced by joint researchers in Yale University and Columbia University, and ranked 173 out of 180 countries. People are more interested in PM_{10} levels now and there are many articles about fine particulate pollution (PM_{10}) in the media in Korea. PM_{10} is a fine particular with aerodynamic diameter of up to $10\mu m$, which are not filtered by the bronchial tubes and cause many diseases. Shaughnessy *et al.* (2015) reported that the number of patients with lung disease increases as the PM_{10} level increases. Zúñiga *et al.* (2016) found that a high concentration of PM_{10}, ozone and nitrogen dioxide can result in a high mortality from cardiovascular and pulmonary disease.

The Korea Ministry of Environment provides the real-time PM_{10} concentration and next day forecast with four classes: "good" for 0 to 30, "normal" for 31 to 80, "bad" for 81 to 150 and "very bad" for more than 150. Recently, there are many days with PM_{10} classified as bad or very bad in Korea. Therefore, people want to know what causes the high PM_{10} concentration in Korea. Many factors can cause a high PM_{10} concentration; however, most people believe that the major reasons are severe air pollution from China, Korea's thermal energy plants on the west coast, and using old diesel vehicles.

We examine which factors affect the PM_{10} concentration as well as build a forecast model for the next day mean and max of PM_{10} concentration. We believe that this analysis will be helpful to public

[1] Corresponding author: Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. E-mail: josong@ewha.ac.kr

Published 31 March 2018 / journal homepage: http://csam.or.kr

Table 1: Description of variables

Category	Variables	Type	
Air pollutant	fine particular (PM_{10})	hourly	numeric
	ozone (O_3)	daily	numeric
	carbon monoxide (CO)	daily	numeric
	sulfur dioxide (SO_2)	daily	numeric
	nitrogen dioxide (NO_2)	daily	numeric
Meteorological elements	temperature	daily	numeric
	wind speed	daily	numeric
	wind direction	daily	category
	relative humidity	daily	numeric
	sea level pressure	daily	numeric
	hours of daylight	daily	numeric
	duration of fog	daily	numeric
	precipitation	daily	numeric
	duration of precipitation	daily	numeric
	1 hour solar radiation	daily	numeric
	solar radiation	daily	numeric
	fresh snow cover	daily	numeric
	amount of clouds	daily	numeric
	yellow warning	daily	category
China air quality	beijing $PM_{2.5}$	daily	numeric

and policy makers in Korea. There are several studies about PM_{10} forecasting. Sayegh *et al.* (2014) used a multiple regression, GAM and QPM model to forecast PM_{10}. Perez and Reyes (2006) and Taneja *et al.* (2016) used an integrated neural network and SARIMA (Seasonal ARIMA). In addition, Chaloulakou *et al.* (2003), Hooyberghs *et al.* (2005), Nejadkoorki and Baroutian (2012), Poggi and Portier (2011), and Cheng *et al.* (2013) also forecast PM_{10} using a linear regression and clustering method.

In this paper, we will use seven different models to forecast PM_{10} concentrations: two time series models (ARIMA, ARFIMA) and five regression models (linear regression, Randomforest, support vector machine (SVM), boosting, neural network). We will use several explanatory variables in our analysis. They are $PM_{2.5}$ in Beijing, air pollutants, yellow sand and meteorological elements.

The paper is and organized as follows. In Section 2, we explain the data and the variables used in the analysis. We explain the preparation of variables in detail because all data are time series data. In Section 3, we compare several PM_{10} forecasting models and find the best model for both regression and classification. Section 4 provides the concluding remarks.

2. Data

In this section, we describe the dataset used in forecasting daily PM_{10} and preparation of variables.

2.1. Data collection

We consider the various explanatory variables for forecasting daily PM_{10} concentration. In Table 1 provides the descriptions of and types of variables. Data were collected from 2011/08/01 to 2015/07/31.

PM_{10} level is recorded hourly; however, we will use a daily PM_{10} mean and max as the response variables in our analysis. However, these hourly PM_{10} values can be used as explanatory variables in the model. Air pollution data are obtained from Air Korea (www.airkorea.or.kr/realSearch), which indicates real-time air pollution levels recorded by the Korea Environmental Management Corporation.

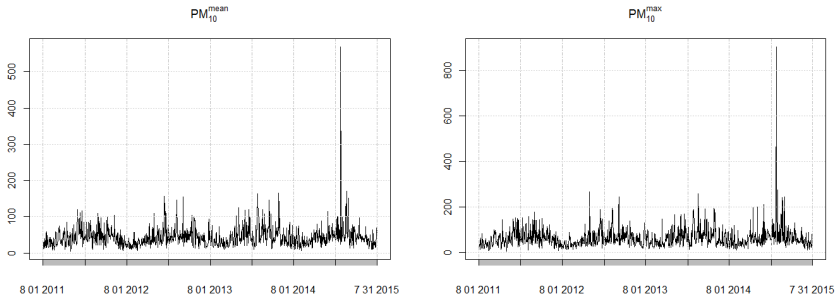

Figure 1: *Time series plots of PM_{10}^{mean} and PM_{10}^{max}.*

Meteorological element information is obtained from weather information portal (data.kma.go.kr) and yellow dust information is obtained from the Korea Meteorological Administration. Beijing $PM_{2.5}$ is collected from StateAir (Air Quality Monitoring Program), which is managed by U.S. Department of State.

2.2. Missing data

Some variables contain missing values. For meteorological data, precipitation, maximum fresh snow cover, and duration of fog are missing values not offered by the Meteorological Administration, if there was no snow, rain and fog. These missing values are replaced by 0. Sea-level pressure and hours of daylight have one missing value and 1-hour solar radiation and solar radiation have 13 missing values. These missing values are imputed by K-nearest neighbor (KNN) method which replaces the missing value using only K-nearest observations.

2.3. Data preparation

Our goal of study is to forecast daily mean and max PM_{10} as well as class of PM_{10}, using variables that are expected to affect the concentration of PM_{10}. In this section, we describe how we build each explanatory variable in our model. Let x_t be a vector of data x at time t, $t = 1, \ldots, n$ and n be the number of data. For the time series variables, we consider the difference series $\Delta X_t = X_t - X_{t-1}$. We also consider the explanatory variable up to lag p as $\{x_{t-1}, \ldots, x_{t-p}\}$ because the current PM_{10} level is affected by previous explanatory variables and use previous data to forecast future PM_{10} level. This lag p can be different for each variable and we select proper p using autocorrelation function (ACF) of original series and difference series. We denote the mean and max of PM_{10} as PM_{10}^{mean} and PM_{10}^{max}. More detailed procedures are as follows.

2.3.1. PM_{10} concentration

Figure 1 displays time series plot of daily mean and max PM_{10} concentration in Seoul from 2011/08/01 to 2015/07/31. Both of PM_{10}^{mean} and PM_{10}^{max} show similar trends by year. We can see that PM_{10}^{mean} and PM_{10}^{max} levels are high in the winter and spring, and low in the summer and fall. PM_{10} seems exceptionally high in February 2015. In year 2015, PM_{10}^{mean} is 568.6 $\mu g/m^3$ on February 23, the PM_{10}^{max} is 880.36 $\mu g/m^3$ on February 8 and 901.52 $\mu g/m^3$ on February 23 when the yellow dust warning was issued.

The previous PM_{10} is expected to have influence on the present PM_{10}; therefore, we consider previous PM_{10} as explanatory variables. A period of past PM_{10} affecting current PM_{10} is determined

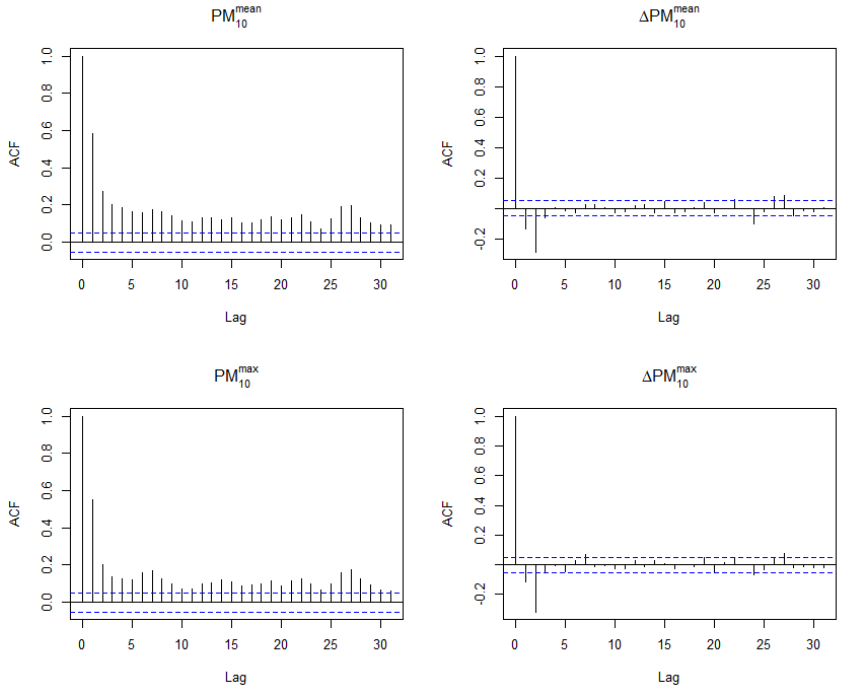

Figure 2: *ACF plots of original data and differential data. ACF = autocorrelation function.*

by examining the ACF plot. Figure 2 is the ACF plots of original series and difference series for PM_{10}^{mean} and PM_{10}^{max}. We can see that the original series have long memory. Therefore, we have to include too many explanatory variables in our model if we consider all significant lags for ACF. On the contrary, the differential data has short memory and it is possible to consider only a few past PM_{10} levels as explanatory variables. Thus, we forecast differential PM_{10}^{mean} and PM_{10}^{max} in our model. The differential PM_{10}^{mean} and PM_{10}^{max} is denoted by ΔPM_{10}^{mean} and ΔPM_{10}^{max}.

Previous ΔPM_{10}^{mean} and ΔPM_{10}^{max} will be used as explanatory variables and we have to select a proper log p in our model. We select the appropriate lag of ΔPM_{10}^{mean} and ΔPM_{10}^{max} based on the performance of the ARIMA model for one-step out of forecast. We partition data from 2011/08/01 to 2013/07/31 as a training set and data from 2013/08/01 to 2014/07/31 as a test set. As the performance measures, we consider root mean squared error (RMSE) and mean absolute error (MAE):

$$\text{RMSE} = \frac{1}{n}\sqrt{\sum_{t=t_0+1}^{n}\left(\Delta \text{PM}_{10,t} - \widehat{\Delta \text{PM}}_{10,t}\right)^2},$$

$$\text{MAE} = \frac{1}{n}\sum_{t=t_0+1}^{n}\left|\Delta \text{PM}_{10,t} - \widehat{\Delta \text{PM}}_{10,t}\right|,$$

where t_0 is the number of train data and $\widehat{\text{PM}}_{10,t}$ is one-step forecast PM_{10} at time t. Figure 3 provides RMSE and MAE plots based on ARIMA $(p, 1, 0)$ model. The decrement of RMSE and MAE for ΔPM_{10}^{mean} and ΔPM_{10}^{max} are large in lag up to three, after that, the decrement seems slight. Therefore,

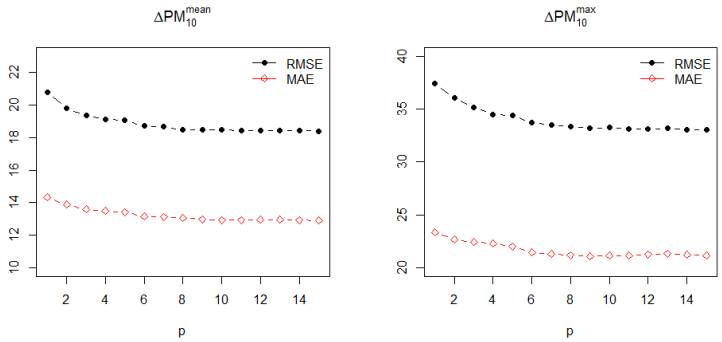

Figure 3: *RMSE and MAE plots for ARIMA(p) models. RMSE = root mean squared error; MAE = mean absolute error.*

we consider $\Delta PM_{10,t-1}$, $\Delta PM_{10,t-2}$, $\Delta PM_{10,t-1}$ as explanatory variables to forecast $\Delta PM_{10,t}$. We also include previous PM_{10} up to lag 3 as explanatory variables.

PM_{10} levels from the previous day will also affect mostly next day PM_{10}. Therefore, we investigate more for hourly PM_{10} levels in the previous day. We believe including 24 hourly PM_{10} levels in the model is not appropriate. Hence, we try to find the best time intervals for the previous hourly PM_{10} in our model. We used tree models and various correlation plots. We found that these four intervals are easy to interpret and one of the best means to forecast next day PM_{10} levels. We denote that $PM_{10,t-1}^{mean,1}$ is the mean PM_{10} value for 0 to 6 hour in the previous day; $PM_{10}^{mean,2}$ is for 6 to 12; $PM_{10}^{mean,3}$ is for 12 to 18; $PM_{10}^{mean,4}$ is for 18 to 24.

We also consider that the day and month effect for PM_{10}. PM_{10} concentration is anticipated to be high in the spring due to yellow dust and on week days due to commuter transport in Seoul. In order to confirm that PM_{10} actually differs by month and day, we made box plots of PM_{10}^{mean} and PM_{10}^{max} by month and day in Figure 4. We remove outliers in order to see the difference clearly. The outliers exist on February 23 for both of PM_{10}^{mean} and PM_{10}^{max} and on February 22 only for PM_{10}^{max}. The figure of PM_{10}^{mean} by month reveals that PM_{10}^{mean} is higher in November–May than in June–October. The daily PM_{10}^{mean} is higher on weekdays than weekends. Dummy variables for months and days are included in the explanatory variables.

Consequently, the explanatory variables for $\Delta PM_{10,t}$ are determined as past 3 days PM_{10} ($PM_{10,t-1}$, $PM_{10,t-2}$, $PM_{10,t-3}$), past 3 days difference PM_{10} ($\Delta PM_{10,t-1}$, $\Delta PM_{10,t-2}$, $\Delta PM_{10,t-3}$), hourly PM_{10} on previous day ($PM_{10,t-1}^{1}, PM_{10,t-1}^{2}, PM_{10,t-1}^{3}, PM_{10,t-1}^{4}$) and dummy variables for months and days.

2.3.2. Air pollution concentration

Air pollutants are commonly known as ozone (O_3), carbon monoxide (CO), sulfur dioxide (SO_2) and nitrogen dioxide (NO_2). Carbon monoxide is known to be generated by the incomplete combustion of carbon which is mainly emitted by transportation methods. Sulfur dioxide is released when fossil fuels, such as coal and petroleum their contains sulfur, are burned. It is mainly generated in power plants, heating equipment, and industrial processes. Nitrogen dioxide is generated by the oxidation of nitrogen monoxide which is made from vehicle exhaust and power plants. We consider air pollutants as explanatory variables since the increase in automobile emissions and coal consumption is one of the main causes of PM_{10}.

We can use daily mean or max or both as explanatory variables in our model. We examined the

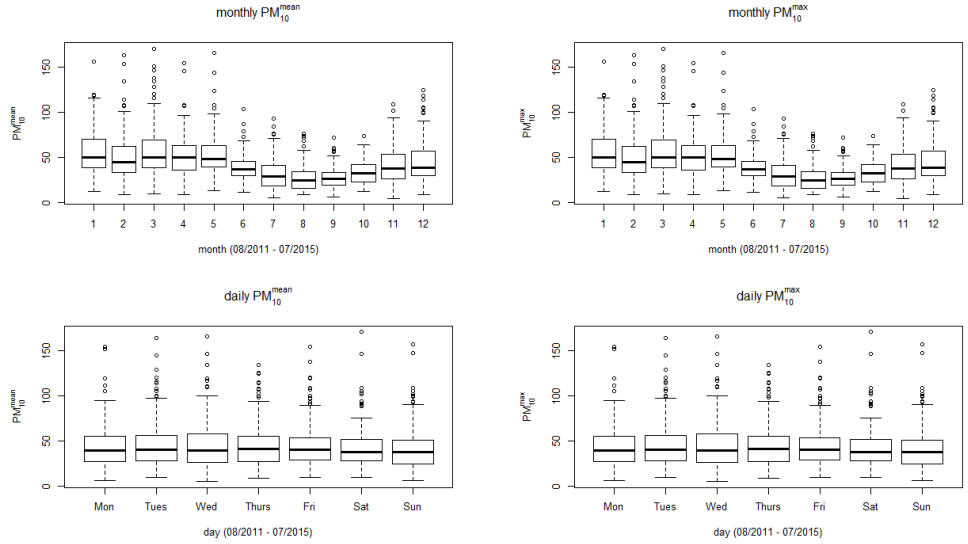

Figure 4: *Box plots of PM_{10}^{mean} and PM_{10}^{max} by month and day.*

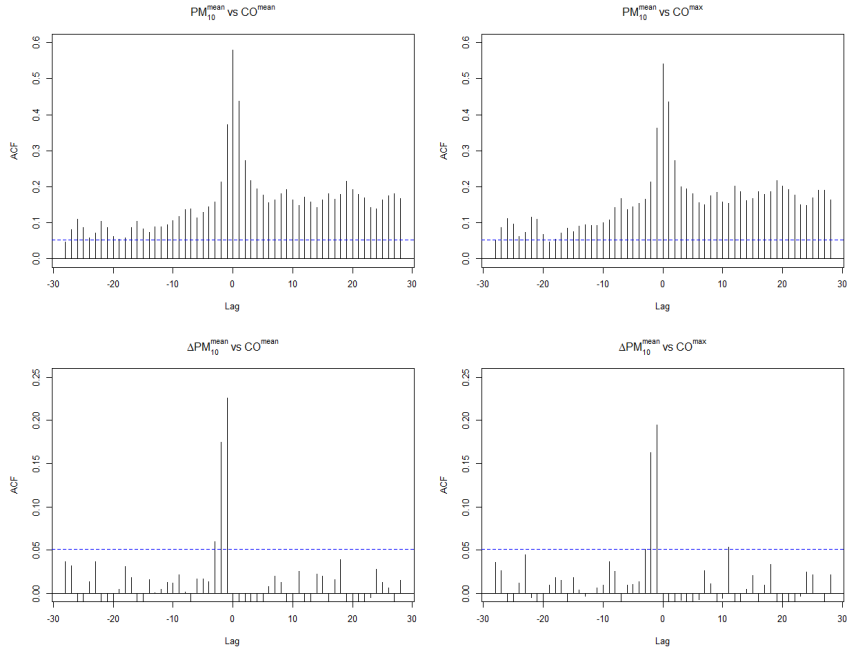

Figure 5: *CCF plots for PM_{10}^{mean} and ΔPM_{10}^{mean}. CCF = cross and covariance and correlation function.*

relation between these variables and PM_{10}^{mean} from four plots in Figure 5. The two plots in top row are cross and covariance and correlation function (CCF) of PM_{10}^{mean} and CO^{mean}, and CCF of PM_{10}^{mean} and CO^{max}. Two plots in bottom row are CCF of ΔPM_{10}^{mean} and CO^{mean}, and CCF of ΔPM_{10}^{mean} and CO^{max}. In CCF plots, PM_{10}^{mean} and CO^{mean} or CO^{max} have long memory, but ΔPM_{10}^{mean} and CO^{mean}

Table 2: Description of response and selected explanatory variables

Category	Variables	Variable explanation	Variables	Variable explanation
Response	$\Delta PM^{mean}_{10,t}$	difference daily mean PM_{10} (present - 1day ago)	$\Delta PM^{max}_{10,t}$	difference daily max PM_{10} (present–1day ago)
Previous PM_{10}	$PM^{mean}_{10,t-1}$	daily mean PM_{10} (1 day ago)	$PM^{max}_{10,t-1}$	daily max PM_{10} (1 day ago)
	$PM^{mean}_{10,t-2}$	daily mean PM_{10} (2 day ago)	$PM^{max}_{10,t-2}$	daily max PM_{10} (2 day ago)
	$PM^{mean}_{10,t-1}$	daily mean PM_{10} (3 day ago)	$PM^{max}_{10,t-3}$	daily max PM_{10} (3 day ago)
	$\Delta PM^{mean}_{10,t-1}$	difference daily mean PM_{10} (1 day–2 day ago)	$\Delta PM^{max}_{10,t-1}$	difference daily max PM_{10} (1 day–2 day ago)
	$\Delta PM^{mean}_{10,t-2}$	difference daily mean PM_{10} (2 day–3 day ago)	$\Delta PM^{max}_{10,t-2}$	difference daily max PM_{10} (2 day–3 day ago)
	$\Delta PM^{mean}_{10,t-1}$	difference daily mean PM_{10} (3 day–4 day ago)	$\Delta PM^{max}_{10,t-3}$	difference daily max PM_{10} (3 day–4 day ago)
	$PM^{mean,1}_{10,t-1}$	mean PM_{10} 0–6 hours	$PM^{max,1}_{10,t-1}$	max PM_{10} 0–6 hours
	$PM^{mean,2}_{10,t-1}$	mean PM_{10} 6–12 hours	$PM^{max,2}_{10,t-1}$	max PM_{10} 6–12 hours
	$PM^{mean,3}_{10,t-1}$	mean PM_{10} 12–18 hours	$PM^{max,3}_{10,t-1}$	max PM_{10} 12–18 hours
	$PM^{mean,4}_{10,t-1}$	mean PM_{10} 18–24 hours	$PM^{max,4}_{10,t-1}$	max PM_{10} 18–24 hours
	month.int	month (January–December)	month.int	month (January–December)
	day.int	Mon, Tues, Wed, Thurs, Fri, Sat, Sun	day.int	Mon, Tues, Wed, Thurs, Fri, Sat, Sun
Air pollutant	$SO^{mean}_{2,t-1}$	mean SO_2 (1 day ago)	$SO^{mean}_{2,t-1}$	mean SO_2 (1 day ago)
	$SO^{mean}_{2,t-2}$	mean SO_2 (2 day ago)	$SO^{mean}_{2,t-2}$	mean SO_2 (2 day ago)
	CO^{mean}_{t-1}	mean CO (1 day ago)	CO^{mean}_{t-1}	mean CO (1 day ago)
	CO^{mean}_{t-2}	mean CO (2 day ago)	CO^{mean}_{t-2}	mean CO (2 day ago)
	$NO^{mean}_{2,t-1}$	mean NO_2 (1 day ago)	$NO^{mean}_{2,t-1}$	mean NO_2 (1 day ago)
	$NO^{mean}_{2,t-2}$	mean NO_2 (2 day ago)	$NO^{mean}_{2,t-2}$	mean NO_2 (2 day ago)
	$O^{mean}_{3,t-1}$	mean O_3 (1 day ago)	$O^{mean}_{3,t-1}$	mean O_3 (1 day ago)
Meteorological elements	tem^{mean}_{t-1}	mean temperature (1 day ago)	tem^{mean}_{t-1}	mean temperature (1 day ago)
	$speed^{mean}_{t-1}$	mean wind speed (1 day ago)	$speed^{mean}_{t-1}$	mean wind speed (1 day ago)
	dir_{t-1}	wind direction (16 bearing) (1 day ago)	dir_{t-1}	wind direction (16 bearing) (1 day ago)
	dir_{t-2}	wind direction (16 bearing) (2 day ago)	$humid^{mean}_{t-1}$	mean relative humidity (%) (1 day ago)
	$humid^{mean}_{t-1}$	mean relative humidity (%) (1 day ago)	$humid^{mean}_{t-2}$	mean relative humidity (%) (2 day ago)
	$rain.hour_{t-1}$	duration of precipitation (1 day ago)	$rain.hour_{t-1}$	duration of precipitation (1 day ago)
	$rain_{t-1}$	precipitation (1 day ago)	$rain_{t-1}$	precipitation (1 day ago)
	$press^{mean}_{t-1}$	mean sea level pressure (1 day ago)	$press^{mean}_{t-1}$	mean sea level pressure (1 day ago)
	$sun.sum_{t-1}$	solar radiation (1 day ago)	$sun.sum_{t-1}$	solar radiation (1 day ago)
	$sun.hour_{t-1}$	hours of daylight (1 day ago)	$sun.hour_{t-1}$	hours of daylight (1 day ago)
	$cloud.hour_{t-1}$	mean amount of clouds (1 day ago)	$cloud.hour_{t-1}$	mean amount of clouds (1 day ago)
	$fog.hour_{t-1}$	duration of fog (1 day ago)	$fog.hour_{t-1}$	duration of fog (1 day ago)
	$sun.high_{t-1}$	1 hour solar radiation (1 day ago)	$sun.high_{t-1}$	1 hour solar radiation (1 day ago)
	$dust_{t-1}$	yellow dust warning	$dust_{t-1}$	yellow dust warning
China air quality	$beijing^{mean}_{t-1}$	beijing $PM_{2.5}$ (1 day ago)	$beijing^{mean}_{t-1}$	beijing $PM_{2.5}$ (1 day ago)
	$beijing^{mean}_{t-2}$	beijing $PM_{2.5}$ (2 day ago)	$beijing^{mean}_{t-2}$	beijing $PM_{2.5}$ (2 day ago)
	$beijing^{mean}_{t-3}$	beijing $PM_{2.5}$ (3 day ago)	$beijing^{mean}_{t-3}$	beijing $PM_{2.5}$ (3 day ago)

or CO^{max} have short memory. Accordingly, if we use ΔPM_{10} as response variables, we can consider only a few explanatory variables by selecting the variables up to significant lags in CCF plots. We can see that CO^{mean} and CO^{max} have a similar pattern. However, CO^{mean} has a higher correlation without ΔPM^{mean}_{10}. Therefore, we will use CO^{mean} in our model only up to lag = 2. For the other air pollutant variables, we repeat the above procedure and then select the explanatory variables. We also did the same procedures for PM^{max}_{10}. Table 2 presents the selected air pollutants variables.

2.3.3. Meteorological elements

Air quality is influenced by wind, humidity, duration of fog, precipitation and yellow dust. Among the meteorological elements, yellow dust is a phenomenon characterized by finest dust that originates in China and Inter Mongolia, that blows in on the westerlies. A yellow dust watch is issued if the level is greater than 800 and expected to last more than 2 hours. It is likely that the yellow dust level affects PM_{10} level. Therefore we include this variable in our model: if the yellow dust watch is issued in the previous day, the variable is one; otherwise, the variable is zero. The other meteorological elements are also considered as explanatory variables and these time series variables are selected thorough the same method in Section 2.3.2. Table 2 presents the selected weather-related variables.

2.3.4. $PM_{2.5}$ concentration in Beijing

The concentration of PM_{10} in Korea is conjectured to be affected by the air quality in China due to its geographical proximity. In order to address the influence, we consider the $PM_{2.5}$ in Beijing which is the measure of level of fine dust in China. The explanatory variables related to $PM_{2.5}$ were selected by the same method as in Section 2.3.2; Table 2 presents the selected variables. Finally, Table 2 shows the explanatory variables used for forecasting ΔPM_{10}^{mean} and ΔPM_{10}^{max}.

3. Analysis

This section explains how to construct an optimal forecasting model for PM_{10}^{mean} and PM_{10}^{max}. We will also explain the classification model to forecast four classes of PM_{10}. We use four years of data (2011/08/01–2015/07/31) in our analysis. The last one year of data (2014/08/01–2015/07/31) will be used as a test data to find the optimal model. We believe that the optimal model is the model with the best forecast performance in this test data. For the measure of the performance, we will use as the RMSE for regression and the misclassification rate for classification. The model with the smallest RMSE or the misclassification rate in the test data is the best model.

We have to consider several factors in order to find the best model: the periods of time and window types in a training data. We can use either 1 or 2 or 3 years of data to fit a model in a training set since we have 3 years of data in a training set. We can also use either a growing window or moving window type. A more detailed explanation follows. Assume that we have a dataset up to time $t+1$ and we want to forecast ΔPM_{10} at time $t+2$. A growing window is using the data from 1 to $t+1$ time for ΔPM_{10} forecasts. Whereas, moving window is using the data from 2 to $t+1$ time, if we set to window size as t. Therefore, a growing window is the forecast method using all of the train data and moving window is the method that uses data within a certain period of time. The comparison is also conducted for seven forecast models: two time series models such as ARIMA proposed by Box and Jenkins (1976) and ARFIMA proposed by Granger and Roselyne (1980); five regression models as linear regression with stepwise procedure, Randomforest (Breiman, 2001), Gradient boosting model (Friedman, 2002; Ridgeway, 2012), Neural Network (Hastie *et al.*, 2009) and SVM (Corte and Vapnik, 1995). The seven models are briefly described in Park *et al.* (2011) and Hastie *et al.* (2009). Therefore we consider 42 combinations: three time periods (1, 2, 3 years), two window types (growing, moving window) and seven regression models.

3.1. Forecasting PM_{10} in Seoul

We forecast daily mean PM_{10} in Section 3.1.1 and daily max PM_{10} in Section 3.1.2. We forecast ΔPM_{10} first and then convert ΔPM_{10} into PM_{10}. We explain how to find the optimal model in detail. Since we have four years of data (2011/08/01–2015/07/31). We use the first three years of data as a training set and the last one year of data as a test set. The best model is the model with a minimum RMSE in the test set. We need to find the optimal tuning parameter for each method using only training data. Therefore, we partition training data according to time periods. For example, if we use 1 year time period, we use 1 year of data (2011/08/01–2012/07/31) as a training set and the last 2 years of data (2012/08/01–2014/07/31) as a test set. If we use the 2 year time period, we first use 2 year of data as a training and the last 1 year of data as a test. We cannot partition the training data for the 3 year time period since we have only 3 years of data. Therefore, we use the optimal tuning parameter values from the 2 year time period model (Figure 6).

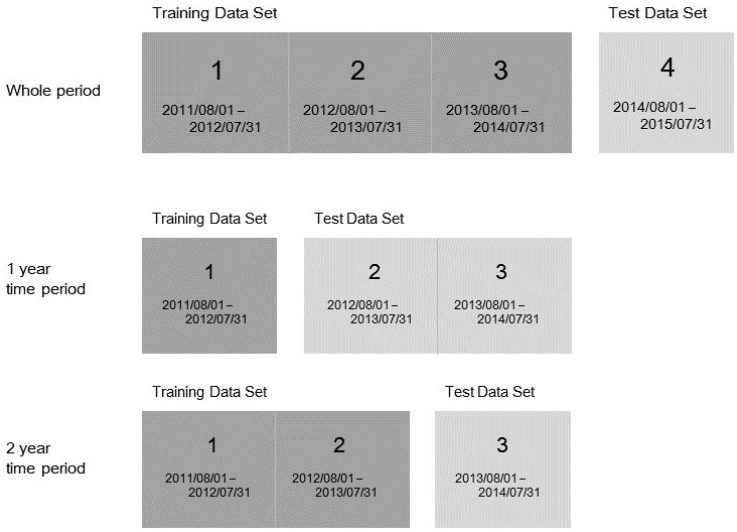

Figure 6: *Explanation of training data set and test data set.*

Table 3: Set of explanatory variables related to previous PM_{10}

	Previous daily PM_{10}	Previous differential PM_{10}	Previous hourly PM_{10}
Set 1	$PM_{10,t-1}$, $PM_{10,t-2}$, $PM_{10,t-3}$	$\Delta PM_{10,t-1}$, $\Delta PM_{10,t-2}$, $\Delta PM_{10,t-3}$	$PM^1_{10,t-1}$, $PM^2_{10,t-1}$, $PM^3_{10,t-1}$, $PM^4_{10,t-1}$
Set 2		$\Delta PM_{10,t-1}$, $\Delta PM_{10,t-2}$, $\Delta PM_{10,t-3}$	$PM^1_{10,t-1}$, $PM^2_{10,t-1}$, $PM^3_{10,t-1}$, $PM^4_{10,t-1}$
Set 3			$PM^1_{10,t-1}$, $PM^2_{10,t-1}$, $PM^3_{10,t-1}$, $PM^4_{10,t-1}$

3.1.1. Daily mean PM_{10}

We forecast mean PM_{10} using the seven models with the explanatory variables in Table 2 and compare forecast performances for the seven models. For each model, we tried to find the optimal tuning parameters with the above procedure. For example, for the 2 year training set, we obtained the optimal tuning parameters as: for ARIMA, $(p, d, q) = (3, 1, 0)$; for ARFIMA, $(p, d, q) = (10, 0.47, 10)$; for SVM, (cost, kernel) = (0.09, linear); for Neural Network, (size, decay) = (4, 25); for Boosting, (shrink, ntree, interaction depth) = $(0.01, 900, 4)$; for Randomforest, mtry = 7.

We cannot use all possible regression methods to find the optimal model since there are many of explanatory variables. We use a stepwise regression for variable selection; however, we found that including all available variables in the model does not necessarily provide the best result. Therefore, we tried three different sets of PM_{10} related variables described in Table 3. You can see that set 1 has all PM_{10} related variables. Set 2 does not have previous daily PM_{10} variables. Set 3 has only previous hourly PM_{10} variables. Consequently, Set 3 gives the best forecast performance and we report the result using only Set 3 for PM_{10} related variables found in Table 4.

RMSEs are not significantly different by time periods and window types. However, it shows large differences among forecast models. Especially, the regression models have a better forecast performance than time series models. This indicates that the forecast performance improves by including explanatory variables. Among the seven models, linear regression and SVM performs better than other models. The best model is the linear regression model with a 2 year time period and a moving window type.

Table 4: Test error of the models for PM_{10}^{mean}

	1 year		2 year		3 year	
	Growing	Moving	Growing	Moving	Growing	Moving
Linear regression	**18.90**	19.81	18.92	**18.82**	19.23	**19.11**
ARIMA	37.15	39.11	35.49	36.15	34.69	34.91
ARFIMA	38.28	34.80	35.86	36.50	34.62	40.72
Randomforest	36.14	36.59	35.88	36.20	36.19	36.14
Boosting	33.94	34.51	33.78	33.47	33.93	34.05
Neural Network	37.68	37.67	37.67	37.67	37.67	37.67
SVM	19.20	19.68	19.44	19.26	19.30	19.51

Note: Bold type is the smallest test error for each period of the train data. SVM = support vector machine.

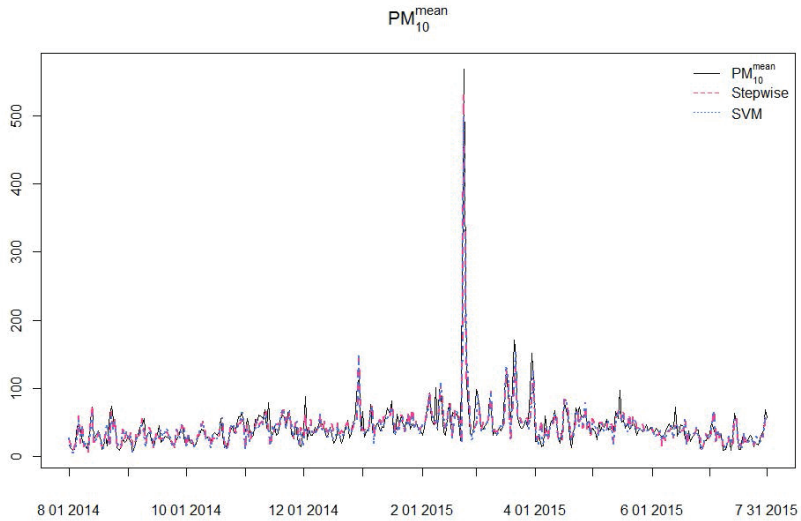

Figure 7: *Time series plot of PM_{10}^{mean} forecasts from linear regression and SVM. SVM = support vector machine.*

Table 5: Important variables in forecasting PM_{10}^{mean}

Category	Increase (+)	Decrease (−)
meteorological elements	$humid_{t-1}$, $sun.hour_{t-1}$	$cloud.mean_{t-1}$, tem_{t-1}, $press_{t-1}$
air pollutant	$NO_{2,t-1}$	
month	May	January–April, June–December
day	weekdays	weekend
hourly mean PM_{10}	18–24 hour	0–18 hour
China air quality	$beijing_{t-1}$	$beijing_{t-2}$

Figure 7 displays time series plot of forecast from linear regression and SVM. Both models forecast PM_{10} adequately. The performance difference between linear regression and SVM is very small; however, we choose the linear regression model as the best model better interpretability for daily PM_{10}^{mean} because it has.

Table 5 represents the sign of the selected variables from the final (best) linear regression model. The table shows that among the meteorological variables on the previous day, humidity, cloudiness, temperature, sea level pressure and hours of daylight influence daily mean PM_{10} on the present day. The daily mean PM_{10} increases as the humidity and hours of daylight increases; however, the increase in other meteorological factors leads to an decrease in daily mean PM_{10}. Among air pollutants, only

Table 6: Test error of the models for PM_{10}^{max}

	1 year		2 year		3 year	
	Growing	Moving	Growing	Moving	Growing	Moving
Linear regression	**52.68**	54.05	**52.48**	52.54	**51.97**	52.05
ARIMA	63.45	65.16	62.39	62.49	61.75	62.20
ARFIMA	64.49	62.32	62.11	63.02	61.96	67.26
Randomforest	64.21	64.43	63.78	64.10	63.17	63.33
Boosting	63.73	64.79	62.84	63.31	62.59	62.47
Neural Network	69.06	69.06	69.08	69.06	69.05	69.06
SVM	52.95	53.22	52.98	52.82	52.73	52.65

Note: Bold type is the smallest test error for each period of the train data. SVM = support vector machine.

sulfur dioxide affects mean PM_{10}. Sulfur dioxide is a by-product from power plants or heating equipment; therefore, we can see the relationship between PM_{10}^{mean} and these facilities. The mean PM_{10} is also more in May than other months. The only month with a positive coefficient is May. All other months have negative coefficients. However, if we examine the coefficient values, we can see that the coefficient for summer and autumn are lower than winter and spring. It coincides with the boxplot of monthly PM_{10} in Figure 4. The mean PM_{10} also increases more weekdays than on weekends. The mean PM_{10} increases on the present day as the mean PM_{10} for 18 to 24 hour in the previous day increases. The mean PM_{10} for 0 to 18 hour shows the opposite. The increase of Beijing $PM_{2.5}$, one of the variables of interest, leads to PM_{10}^{mean} increasing on the present day. But, the increase of Beijing $PM_{2.5}$ on the day before yesterday shows the opposite.

3.1.2. Daily max PM_{10}

In the same way as Section 3.1.1, we forecast max PM_{10} with the proposed seven models. We tried to find the optimal tuning parameters for each model. For example, for 2 year training set, we found $(p, d, q) = (3, 1, 0)$ for ARIMA; $(p, d, q) = (3, 0.38, 7)$ for ARFIMA; (cost, kernel) = (0.1, linear) for SVM; (size, decay) = (16, 10) for Neural Network; (shrink, ntree, interaction depth) = $(0.01, 2000, 2)$ for Boosting; mtry = 16 for Randomforest. We also consider three sets of explanatory variables presented in Table 3. Among the three sets, Set 3 has the best forecast performance as before (Table 6).

We can see that forecast performance does not depend on time periods or window type. However, the performance is greatly different by the models, similarly in PM_{10}^{mean}. Among the seven models, linear regression and SVM provide better forecast performance than other models. Especially, linear regression with 3 year time period and a growing window has the smallest RMSE.

Figure 8 shows time series plot of forecast PM_{10}^{max} from linear regression and SVM. The trends of the forecast from the two models are very close to the trend of PM_{10}^{max}. Accordingly, the best model for forecasting PM_{10}^{max} is linear regression which provides the obvious relationship between explanatory variables and the response variable. It also has the best forecast performance.

Table 7 provides sign of the selected variables from the final linear regression model. Unlike the PM_{10}^{mean} model, for meteorological elements, duration of precipitation and yellow dust are included in the model. The max PM_{10} increases on the present day as the duration of precipitation decreases on the previous day. However, the max PM_{10} increases on the present day when the yellow dust is occurred on the previous day. For air pollutants, sulfur dioxide and nitrogen dioxide are included as important variables. As sulfur dioxide increases, max PM_{10} increases and the increase of nitrogen dioxide gives the opposite result. These are different with mean PM_{10} model. The max PM_{10} is more increased in weekdays than weekend, and in spring than in summer, autumn and winter. The increase

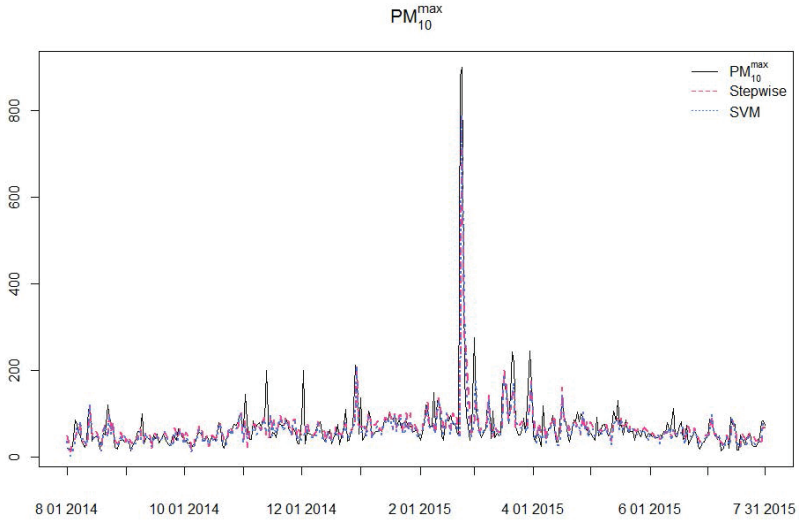

Figure 8: *Time series plot of PM_{10}^{max} forecasts from linear regression and SVM. SVM = support vector machine.*

Table 7: Important variables in forecasting PM_{10}^{max}

Category	Increase (+)	Decrease (−)
meteorological elements	$humid_{t-1}$, $dust_{t-1}$	$cloud_{t-1}^{mean}$, $rain.hour_{t-1}$, $press_{t-1}$
air pollutant	$SO_{2,t-1}$	$NO_{2,t-1}$, $NO_{2,t-2}$
day	Monday, Wednesday–Thursday	Tuesday, Saturday–Sunday
month	March–May	January–February, June–December
hourly max of PM_{10}	18–24 hour	0–18 hour
China air quality	$beijing_{t-1}$	

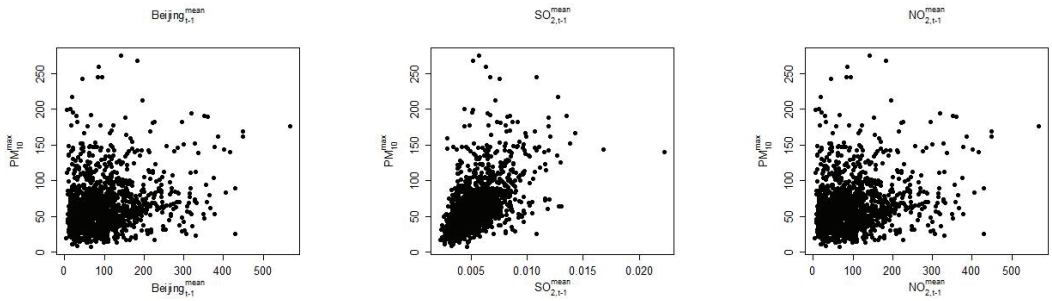

Figure 9: *$PM_{10,t}^{max}$ dependence of important variables ($beijing_{t-1}^{mean}$, $SO_{2,t-1}$, $NO_{2,t-1}$).*

in Beijing $PM_{2.5}$ on the previous day leads to the increase in max PM_{10}.

Figure 9 shows the scatter plots of PM_{10}^{max} and the variables of interest among the significant variables in the linear model. As Beijing $PM_{2.5}$, sulfur dioxide and nitrogen dioxide increase, the max PM_{10} increase. Therefore, we can see that carbon emission from automobiles and ultra fine dust from China have a positive correlation with the max PM_{10}.

Table 8: The frequencies of four categories of PM_{10}^{mean} in the test data

	Good	Normal	Bad	Very bad
2014/08/01–2015/07/31	101	242	18	4

Table 9: Misclassification rate of PM_{10}^{mean} class (Method 1)

	1 year		2 year		3 year	
	Growing	Moving	Growing	Moving	Growing	Moving
Linear regression	0.23	0.24	0.22	0.23	0.22	0.22
ARIMA	0.26	0.27	0.26	0.27	0.26	0.26
ARFIMA	0.29	0.29	0.27	0.27	0.26	0.27
Randomforest	**0.21**	**0.21**	**0.20**	0.21	0.21	0.21
Boosting	0.22	0.22	0.21	0.21	**0.20**	0.21
Neural Network	0.27	0.27	0.27	0.27	0.27	0.27
SVM	0.23	0.22	0.22	0.23	0.24	0.23

Note: bold type is the smallest test error for each period of the train data. SVM = support vector machine.

3.2. Forecasting class of PM_{10} in Seoul

The Korea Ministry of Environment classifies the PM_{10} concentrations by four classes: "good" (0–30), "normal" (31–80), "bad" (81–150) and "very bad" (more than 150). We consider classification models based on these four categories. We can use three different approaches for this classification. The first method (Method 1) is the forecast method using regression models. We forecast PM_{10} using regression models from Section 3.1.1 and then we classify PM_{10} forecasts into four categories: if the predicted value is between 0 to 30, it is classified as "good"; if it is between 31 to 80, it is "normal"; if it is between 81 to 150, it is "bad"; if it is more than 151, it is "very bad". The second method (Method 2) uses numeric class labels as a response. We label the response as 1 if it is "good", 2 if it is "normal", 3 if it is "bad", and 4 if it is "very bad". Then we fit the data using the best regression model in Section 3.1.1. Finally, we classify PM_{10} forecasts according to predicted values: if the forecast is smaller than 1.5, it is "good", if it is between 1.5 and 2.5, it is "normal", if it is between 2.5 and 3.5, it is "bad" and if it is more than 3.5, it is "very bad". The last method (Method 3) uses the classification algorithms. We label the response with four categories and apply several classification methods: logistic regression, linear discriminant analysis (LDA), Randomforest, and SVM. Sections 3.2.1–3.2.3 presents the forecast results for three methods. For the measure of forecast performance, we use a misclassification rate. We believe the model with the lowest misclassification rate in a test data (2014/08/01–2015/07/31) is the best model. Table 8 gives the frequencies of these four categories in the test data. Most of days in the test data are shown to be either good or normal.

3.2.1. Class of daily mean PM_{10} using regression methods (Method 1)

In this section, we forecast the class of daily mean PM_{10} with the regression methods. Table 9 shows the misclassification rate of PM_{10}^{mean} class forecast. The misclassification rates are less than 0.3 for all of the models. Linear regression, Randomforest, Boosting and SVM have similar and good forecast performances. However, contrary to the result of Section 3.1, Randomforest and Boosting have the smallest misclassification rate among the models. We can also see that there is no difference by the periods and window types of train data. The performances of time series models are worse than that of regression models. However the difference between time series and regression models are smaller than Section 3.1.

Table 10 shows confusion matrices for test data (2014/08/01–2015/07/31) obtained from the best

Table 10: Confusion matrices of PM_{10}^{mean} class (Method 1)

	Linear regression				ARFIMA				Randomforest			
	Good	Normal	Bad	Very bad	Good	Normal	Bad	Very bad	Good	Normal	Bad	Very bad
Good	63	38	0	0	53	48	0	0	60	41	0	0
Normal	25	213	4	0	29	202	10	1	15	224	3	0
Bad	0	10	8	0	1	12	4	1	0	11	6	1
Very bad	1	0	2	1	0	2	1	1	1	0	2	1
	Boosting				Neural Network				SVM			
	Good	Normal	Bad	Very bad	Good	Normal	Bad	Very bad	Good	Normal	Bad	Very bad
Good	63	38	0	0	62	37	2	0	68	33	0	0
Normal	19	222	1	0	32	199	11	0	31	207	4	0
Bad	0	10	7	1	0	11	4	3	0	10	7	1
Very bad	1	1	1	1	1	1	1	1	1	0	2	1

SVM = support vector machine.

Table 11: Misclassification rate of PM_{10}^{mean} class (Method 2)

	1 year		2 year		3 year	
	Growing	Moving	Growing	Moving	Growing	Moving
Linear regression	0.22	0.23	0.21	0.21	0.23	0.22
Randomforest	**0.21**	0.22	**0.19**	0.21	**0.19**	0.20
SVM	0.22	0.23	0.24	0.23	0.22	0.23

Note: bold type is the smallest test error for each period of the train data. SVM = support vector machine.

Table 12: Confusion matrices of PM_{10}^{mean} class (Method 2)

	Linear regression				Randomforest				SVM			
	Good	Normal	Bad	Very bad	Good	Normal	Bad	Very bad	Good	Normal	Bad	Very bad
Good	69	32	0	0	64	37	0	0	60	41	0	0
Normal	29	211	2	0	18	223	1	0	23	216	3	0
Bad	1	10	6	1	0	10	8	0	0	10	7	1
Very bad	1	1	1	1	1	1	2	0	1	0	2	1

SVM = support vector machine.

models for each method. Most of misclassifications happen between good and normal. It is because most of observations are in these two categories and are close. The risk of the misclassification is also different by class. It is more risky when bad and very bad are classified as good and normal than the opposite. The number of misclassification for this risky case is 11 for linear regression, 15 for ARFIMA, 12 for Randomforest, 12 for Boosting, 13 for Neural network and 11 for SVM. Again, regression models perform better than time series models.

3.2.2. Class of daily mean PM_{10} with numeric labels (Method 2)

We next forecast the class of daily mean PM_{10} with numeric labels using regression models. We select the explanatory variables again from the proposed method in Section 2 since we cannot use the differential PM_{10}^{mean} as response variable. Linear regression, Randomforest and SVM perform better than other methods; therefore, we consider only these three models in this section.

Table 11 shows the misclassification rates of PM_{10}^{mean} class. The table shows no difference by the periods and window types of train data. Randomforest has the best performance by 0.19. However, the difference for models are not significant. In order to confirm the exact forecast performance for each class, Table 12 provides confusion matrices for each model obtained from the train data showing the best models for each method. The results are very similar to the previous section.

Table 13: Misclassification rate of PM_{10}^{mean} class (Method 3)

	1 year		2 year		3 year	
	Growing	Moving	Growing	Moving	Growing	Moving
Logistic regression	0.31	0.25	0.24	0.23	0.25	0.24
LDA	0.27	0.27	0.27	0.25	0.27	0.25
Randomforest	**0.19**	0.20	**0.21**	**0.21**	0.20	**0.19**
SVM	0.23	0.24	0.23	0.22	0.22	0.23

Note: bold type is the smallest test error for each period of the train data.
LDA = linear discriminant analysis; SVM = support vector machine.

Table 14: Confusion matrices of PM_{10}^{mean} class (Method 3)

	Logistic regression				LDA			
	Good	Normal	Bad	Very bad	Good	Normal	Bad	Very bad
Good	65	35	0	1	61	40	0	0
Normal	32	202	5	3	33	197	12	0
Bad	0	10	6	2	0	10	6	2
Very bad	1	1	0	2	1	1	1	1
	Randomforest				SVM			
	Good	Normal	Bad	Very bad	Good	Normal	Bad	Very bad
Good	66	35	0	0	56	45	0	0
Normal	22	218	2	0	21	220	1	0
Bad	0	11	7	0	0	11	7	0
Very bad	1	2	1	0	1	2	1	0

LDA = linear discriminant analysis; SVM = support vector machine.

3.2.3. Class of daily mean PM_{10} using classification methods (Method 3)

Lastly, we forecast the class of daily mean PM_{10} using classification methods. Table 13 presents the misclassification rate of PM_{10}^{mean} class. For the Method 3, there is slight difference by the period and window type of train data. Randomforest performs the best among the models.

Table 14 is confusion matrices obtained from the best models for each method. The table shows a similar result to the previous section.

In terms of misclassification rate, the performance of Randomforest with Methods 2 and 3 are the same. However, the number of misclassifications for risky cases are different. Randomforest with Method 2 has 12 misclassification and Randomforest with Method 3 has 14 misclassifications for this risky case. We also consider the geometric mean (G-mean) proposed by Kubat *et al.* (1997) as a measure of weighted accuracy since the data set is highly unbalanced for four categories. High G-mean score means a better performance. We relabeled the class "good" and "normal" as "normal" and "bad" and "very bad" as "bad" since the G-mean is defined only when it is two classes. We then calculate the G-mean for several cases. The best two models in terms of G-mean are Randomforest model with numeric labels and classification method. Their G-mean scores are 0.55 and 0.51, respectively. Therefore, our final classification model is the Randomforest model with numeric labels. The important variables selected by the Randomforest model are $PM_{10,t-1}^{mean,3}$, $PM_{10,t-1}^{mean,4}$, $PM_{10,t-1}^{mean}$, and $SO_{2,t-1}^{mean}$.

4. Conclusion

As the concentration of PM_{10} in Korea increases, people are paying more attention to PM_{10} forecasting. Accordingly, we proposed a forecast model for PM_{10} and examined some important features that affect PM_{10} concentration. Since PM_{10} is time series data, we consider explanatory variables that in-

clude the previous PM_{10} and other elements such as air pollutants, meteorological elements and China air quality. In order to determine the optimal lag for these variables, we used several plots including ACF and CCF plots. We consider the significant lags as explanatory variables from the plots of ACF of PM_{10} and plots of CCF of PM_{10} and explanatory variables.

Using these selected variables, we forecast mean and max of PM_{10} with the seven forecast models: two time series models ARIMA, ARFIMA and five regression models Linear regression, SVM, Boosting, Randomforest, Neural Network. We also consider the various training data: three periods of time (1, 2, 3 year) and two window types (growing, moving). We compare forecast performance for 42 combinations (seven model $\times$ six training data) in order to find the optimal forecast model. Among the models, linear regression shows the best forecast performance and makes it possible to interpret obvious relationships between explanatory variables and the response variable. We also found that regression models perform better than pure time series models. However, the forecast performance does not depend on the period and window types of training data.

We investigate the cause of PM_{10} from the selected variables in linear regression. We found that there are seasonal and daily effects on PM_{10} levels. We also found that $PM_{2.5}$ in Beijing and sulfur dioxide related to emissions from power plants affect PM_{10} levels. However, carbon monoxide related to automobile emission is not selected as an important variable in our model. Therefore, we can presume that PM_{10} is more influenced by the air conditions of China and power plants than by automobile emissions. We also find that a dramatic increase of PM_{10} is related to yellow dust.

We next forecast the class of PM_{10}^{mean} with three methods: regression methods (Method 1), regression methods with numeric labels (Method 2) and classification methods (Method 3). All of the methods and models show good forecast performance by having a 0.2 misclassification rate. Among the models, Randomforest has the best forecast performance. Randomforest in Method 2 also shows low risky case which is "bad" and "very bad" are classified as "good" and "normal". Therefore, Randomforest with Method 2 is the best forecast model for PM_{10}^{mean} class.

In order to improve our model, we might consider other explanatory variables such as the daily generation of thermal power plants and daily traffic data. For the classification analysis, we may introduce asymmetric loss to find the best model which has a decision boundary to reflect that the loss for risky cases is larger than less risky cases.

Acknowledgements

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2017R1D1A1B03036078).

References

Box GEP and Jenkins GM (1976). *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco.

Breiman L (2001). Random forests, *Machine Learning*, **45**, 5–32.

Chaloulakou A, Kassomenos P, Spyrellis N, Demokritou P, and Koutrakis P (2003). Measurements of PM_{10} and $PM_{2.5}$ particle concentrations in Athens, Greece, *Atmospheric Environment*, **37**, 649–660.

Cheng S, Wang F, Li J, Chen D, Li M, Zhou Y, and Ren Z (2013). Application of trajectory clustering and source apportionment methods for investigating trans-boundary atmospheric PM_{10} pollution, *Aerosol and Air Quality Research*, **13**, 333–342.

Cortes C and Vapnik V (1995). Support-vector networks, *Machine Learning*, **20**, 273–297.

Friedman JH (2002). Stochastic gradient boosting, *Computational Statistics & Data Analysis*, **38**, 367–378.

Granger CWJ and Roselyne J (1980). An introduction to long-memory time series model and fractional differencing, *Journal of Time Series Analysis*, **1**, 15–29.

Hastie T, Tibshirani R, and Friedman J (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction* (2nd ed), Springer-Verlag, New York.

Hooyberghs J, Mensink C, Dumont G, Fierens F, and Brasseur O (2005). A neural network forecast for daily average PM_{10} concentrations in Belgium, *Atmospheric Environment*, **39**, 3279–3289.

Kubat M, Holte R, and Matwin S (1997). Learning when negative examples abound. In *Proceedings of the 9th European Conference on Machine Learning* (pp. 146–153), Springer, London.

Nejadkoorki F and Baroutian S (2012). Forecasting extreme PM_{10} concentrations using artificial Neural Networks, *International Journal of Environmental Research*, **6**, 277–284.

Park C, Kim Y, Kim J, Song J, and Choi H (2011). *Datamining using R*, Kyowoo, Seoul.

Perez P and Reyes J (2006). An integrated neural network model for PM10 forecasting, *Atmospheric Environment*, **40**, 2845–2851.

Poggi JM and Portier B (2011). PM10 forecasting using clusterwise regression, *Atmospheric Environment*, **45**, 7005–7014.

Ridgeway G (2012). *Generalized Boosted Models: A guide to the gbm package*, Accessed March 31, 2010, from: http://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf

Sayegh AS, Munir S, and Habeebullah TM (2014). Comparing the performance of statistical models for predicting PM_{10} concentrations, *Aerosol and Air Quality Research*, **14**, 653–665.

Shaughnessy WJ, Venigalla MM, and Trump D (2015). Health effects of ambient levels of respirable particulate matter (PM) on healthy, young-adult population, *Atmospheric Environment*, **123**, 102–111.

Taneja K, Ahmad S, Ahmad K, and Attri SD (2016). Time series analysis of aerosol optical depth over New Delhi using Box-Jenkins ARIMA modeling approach, *Atmospheric Pollution Research*, **7**, 585–596.

Zúñiga J, Tarajia M, Herrera V, Urriola W, Gómez B, and Motta J (2016). Assessment of the possible association of air pollutants PM_{10}, O_3, NO_2 with an increase in cardiovascular, respiratory, and diabetes mortality in Panama City, *Medicine*, **95**, e2464.

Received November 2, 2017; Revised December 22, 2017; Accepted December 27, 2017